Alice low-energy AI accelerator for edge applications

Alice 1000 series AI accelerator

- *1 - 40 terra operation per second peak processing capacity depending on HW configuration*
- *Base for low-energy AI solutions in smart devices.*
- *Same architecture, scaled performance and tool support is suitable for both IoT and edge*
- *Open tool chain for application development*
- *Deployable close to data sources*
- *Software defined efficient functionality*

Alice is a massively multicore AI accelerator suitable for edge solutions. It is a good design match for low-energy high-capacity applications in various edge and IoT devices like CPE, ECU, smart sensors and other smart devices. The capacity, on multi terra operation level, targes popular and modern efficient neural networks with large and complex application data.

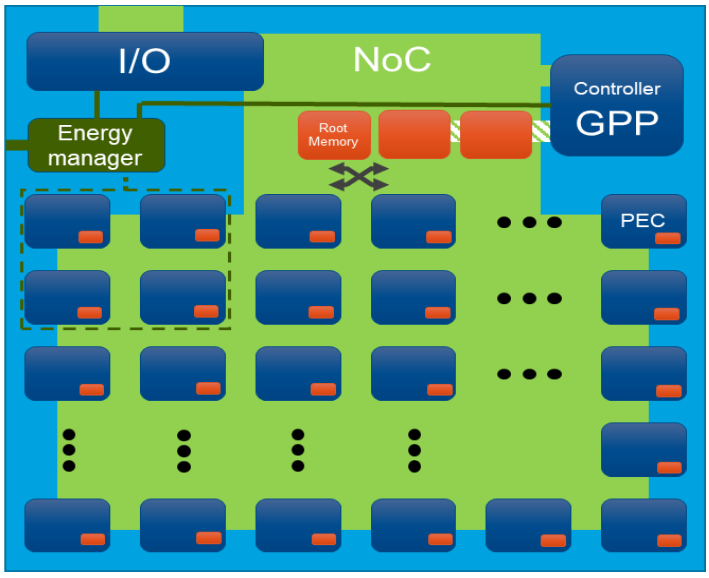
The architecture allows direct control over high performance hardware resources trough micro-programming. The unique flexibility that it provides can be used to extend the range of applications, optimize, and extend the products lifecycle with maintained efficiency.

The chip is qualified for automotive applications and supports security and safety requirements for edge and IoT applications.

The open source and standards-based software and tool support enables modern and fast development processes. The use of standards like ONNX, for AI models, and LLVM-base tools and matching instruction set, helps protecting the customers software and hardware investment and increases platform independence.

Alice key features

- Flexibility trough micro-programming
- High abstraction level programming and automation
- Native and efficient 8-bit integer data processing
- HW support for tensor manipulation
- Managed energy and energy efficiency on cost efficient technology node
- High-utilization AI inference engine
- Whole life cycle tool support
- Open standards
- Evaluation and prototyping support



Alice accelerator architecture

The accelerator is a collection of micro programmed processing elements organized in clusters that are orchestrated from the controlling general-purpose processor (GPP).

Processing Element Cluster (PEC). The processing cluster consist of a memory shared by several micro programmed processing elements (PE). Each is equipped with fast low power vector units containing multiple multipliers. The units can handle 8-bit integer data operations for maximum power efficiency, and it minimizes model parameter size.

Network on Chip (NoC). The PEC are the leaves of the on-chip high-speed star network used for parallel data transport of weights and data between I/O and the local memories in the PEC. The NoC root also contains buffer memory for external and internal data transport. The PEC per2per data transport is handled with a switch.

Controller/GPP. Controls the PECs, manages data transport, handles I/O, the management functions and might host the user application. The GPP is an Imsys IM4000 with a full software stack.

I/O. Contains high speed device interfaces (camera, radar, …) and low speed sensors. Interfaces for control (incl. Ethernet). Memory controllers for DRAM.

Energy manager. Managing sleep, wakeup, clock speed, back-bias, energy surge and other technology dependent and system deployment dependent aspects of keeping the application within the power envelop.

The application will be scheduled on the accelerator array so that the data will be as close to the vector units as possible. Thus, avoiding the cost of memory accesses, long data transport, cashes and MMU.

Benefits

The Alice accelerator concept is a performant AI inference solution deployable in edge, IoT and embedded applications, making the various devices, sensors etc. intelligent.
This will enable scalable and responsive system solutions.

- Scalable, silicon proven, massive manycore technology.
- Customer market differentiation trough software and firmware with preserved efficiency
- Based on Imsys IP
- Also available as IP
- Programmed efficient flexibility trough microprogramming, software and tool support.
- High speed I/O and on chip cache-less data distribution.
- Low energy architecture and circuit design.
- Modular architecture
- Low standby current and fast wakeup
- Secure and functionally safe
- Protects customers software investment. Open standards for LLVM-ISA, OS, Tool-platform and neural network models through the ONNX interchange format.
- Enables scalable system concept

Alice flexibility

- Scalability
  The application is scaled out on the array of PEC by the TVM-based neural network compiler.
- Flexibility
  The PEC's interface is a micro-coded domain-specific ISA and is defined for optimal execution of Neural Networks. The instruction set of the PEC can be extended by micro-programming. The PEC architecture is optimized to support several kernel and stencil-based algorithms, like FFT and matrix operations.
- Energy scaling
  The energy consumption and latency scales close to linear with the array size allocated to the application. Independent of that the latency can be traded against energy consumption with a fivefold energy gain factor.

Alice system features

- Security.
  Transport layer security, TLS, protected control links. Intermediate certificates. Application-defined security on high-speed data links. Signature tests on data can be executed on the accelerator array.
- Management
  The accelerator O&M function covers configuration and fault management. Hardware detection combined with microprogrammed signature detection and management.
- Functional safety
  Builds on fault management capabilities. Lock step enabled. The GPP is a dual core processor. The accelerator array can be used in lock step mode trough microprogramming. Application table – selection of fallback apps. ISO 26262 is built on the basic capabilities of Alice customized to the accelerators system context.
- Energy management
  Managing energy consumption versus performance, hibernation, fast sleep and wakeup, quiescent current, flatten peak current and clock speed.

Software

The GPP is an Imsys IM4000 with a full software stack and tool support. The platform consists of the LLVM based ISA (ISAL), operating system, C-language support and Alice AI accelerator control API.

- The DNN container SW executes the object code from the compiled neural network models and invokes the neural network and digital signal processing operations in the PEC.
- O&M application.

Development Environment – automation

The iterative application design based on Imsys automated tools will enable the AI designer to align the application with the edge requirements. The development tools are designed for rapid turnaround processes, iterative and incremental development. The neural network models are described in the ONNX interchange format that is supported as backend by almost all popular AI design frameworks like Tensorflow, PyTorch …

- Imsys DNN compiler based on the opensource TVM compiler platform with Alice specific optimizations and code generation. These transform the prevalent AI design patterns to an efficient execution schedule.
- The Imsys accelerator simulator lets you verify the application and estimate KPIs for various tentative accelerator configurations.
- The FPGA based Alice emulator is used for evaluation, fast simulation, application prototyping or data collection.

Deployment scenarios and solution context

The accelerator must have access to external memory trough the high speed I/O. The bulky model parameters, weights, are read during execution as well as read and write of temporary data that don't fit the local memories. There is HW support to prepare high speed sensor data for processing. Applications can be hosted on the accelerator's controller.

Imsys AB
Johanneslundsvägen 3
SE-194 61
Upplands Väsby
Sweden
Phone:
+46 8 594 110 70
info@imsystech.com
imsystech.com